苏 州 大 学 实 验 报 告

院、系	计算机学院	年级专业	软件工程	姓名	朱金涛	学号	2327406014
课程名称	机器学习综合实践					成绩	
指导教师	李俊涛	同组实验者	无		实验日期	2025 年 10 月 11 日	

实 验 名 称 机器学习综合实践实验五：朴素贝叶斯分类器

一. 实验目的

1. 掌握文本情感分析的基本原理与实现流程。
2. 学习文本特征提取与机器学习分类模型的应用方法。
3. 理解并验证高斯朴素贝叶斯在文本特征下的失效原因。

二. 实验内容

1. 掌握文本情感分析的基本流程：了解从原始文本到特征向量再到分类器训练的全过程。
2. 熟悉文本预处理与特征提取方法：掌握基于 TF-IDF（词频－逆文档频率）的文本向量化原理与实现。
3. 比较不同分类算法的性能差异：通过 Logistic Regression、Linear SVM 和 Multinomial Naive Bayes 等模型，对比其在情感分类任务上的准确率、精度、召回率和 F1 值。
4. 掌握模型评估与可视化方法：使用混淆矩阵和热力图评估分类结果，直观理解模型误判情况。
5. 分析高斯朴素贝叶斯失效的原因：通过实验验证 GaussianNB 在 TF-IDF 特征上的假设不成立，理解模型分布假设与数据特性的匹配关系。

三. 实验步骤和结果

1. 数据加载与初步分析：

使用 pandas 读取 IMDB 电影评论数据集，检查数据结构。

数据共 25,000 条影评样本，每条影评带有感情标签（0 表示负面，1 表示正面），正负样本数量均衡。

随后将数据按 8:2 比例划分为训练集和测试集，为后续模型训练做准备。

运行结果如下：

```python
X_train, X_test, y_train, y_test =
train_test_split(
    df["text"], df["label"],
    test_size=0.2, random_state=42,
    stratify=df["label"]
)
len(X_train), len(X_test)
```

[8] ✔ 0.0s Python

```
(20000, 5000)
```


教务处制

2. 文本预处理与特征提取：

采用 TF-IDF（词频-逆文档频率）方法将文本转化为数值特征。

处理过程中进行了**小写化、停用词过滤**，并**提取 unigram 与 bigram 特征**。

该方法能突出区分性较强的关键词，使模型在高维稀疏空间中学习到感情差异。

```
tfidf = TfidfVectorizer(
    lowercase=True,              # 全部小写
    stop_words="english",        # 去除常见英文停用词
    ngram_range=(1, 2),          # 使用 unigram + bigram
    min_df=5, max_df=0.9,        # 过滤极少或极常词
    sublinear_tf=True            # 对高频词做对数缩放
)
```

3. 模型训练与性能比较：

分别训练了 Logistic Regression、Linear SVM 和 Multinomial Naive Bayes 三种模型。

为确保公平比较，所有模型均基于相同的 TF-IDF 特征，并在同一训练集上训练。

随后计算各模型在测试集上的准确率、精度、召回率和 F1 值，并汇总成表。

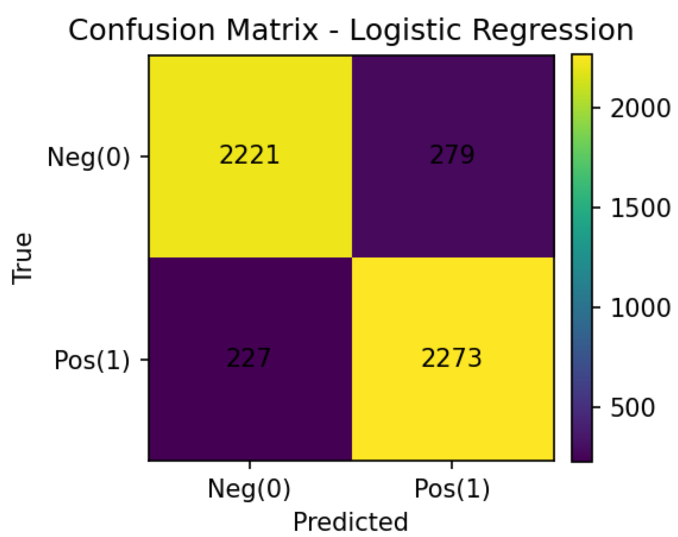
表格如下：

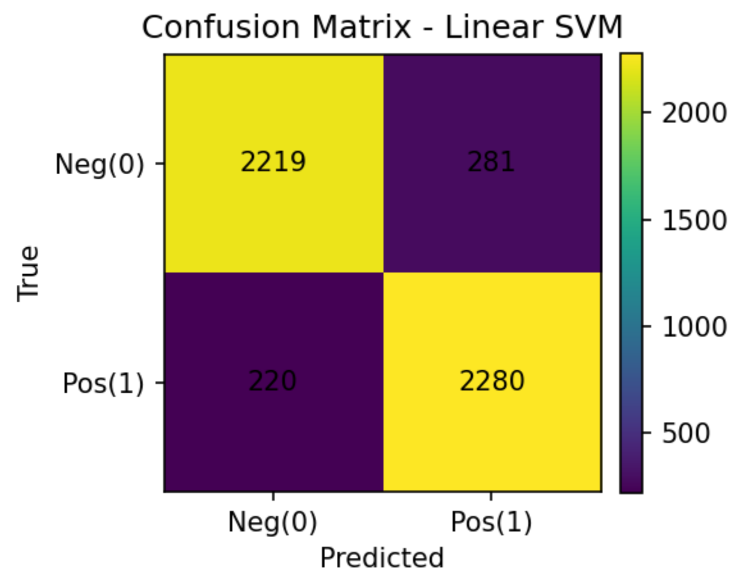
	Accuracy	Precision	Recall	F1
Logistic Regression	0.8988	0.890674	0.9092	0.899842
Linear SVM	0.8998	0.890277	0.9120	0.901008
Multinomial NB	0.8758	0.875050	0.8768	0.875924

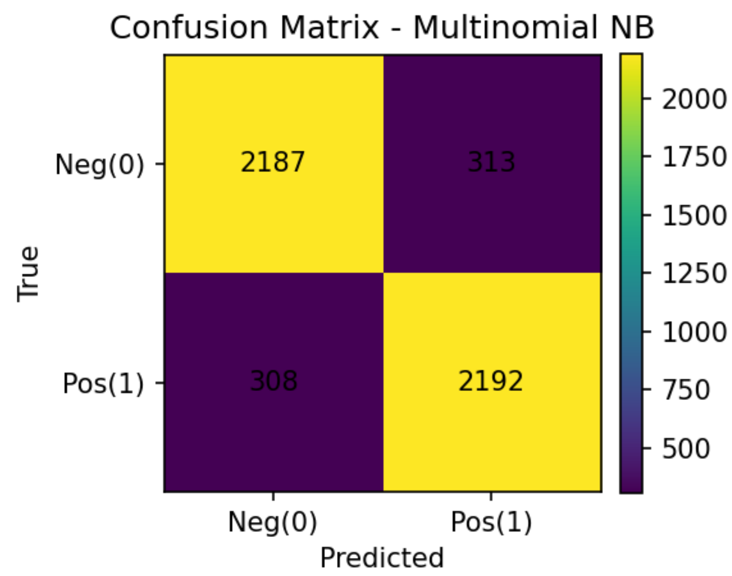
由表格可见，LR 和 LS 模型训练的结果近似且较优，而 MNB 的结果稍差。

4. 混淆矩阵可视化与误差分析：

利用混淆矩阵热力图展示各模型分类结果。







结果显示 Logistic Regression 与 Linear SVM 的正确分类数量较多，误判比例较低；MultinomialNB 的错误主要集中在将部分正面评论误判为负面。

通过热力图分析发现各模型对两类样本的识别较为平衡。

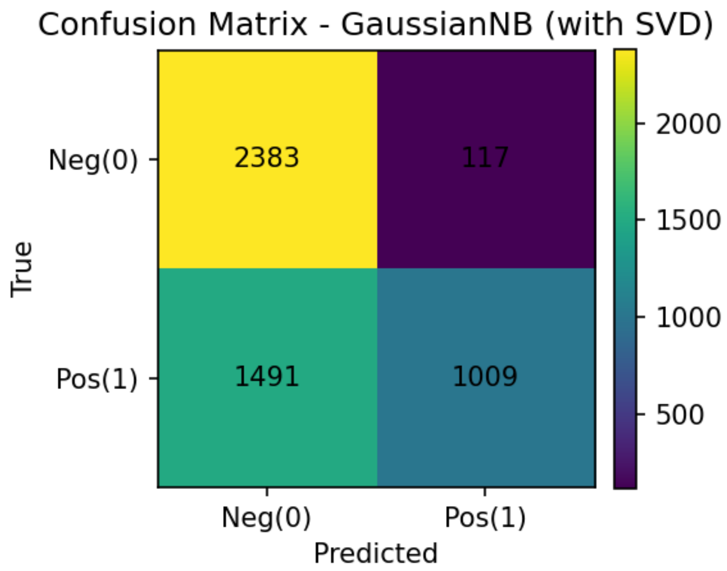
5. 高斯朴素贝叶斯验证：

为验证高斯朴素贝叶斯在文本任务中的适用性，将 TF-IDF 特征经 Truncated SVD 降维后输入 GaussianNB 模型。

运行结果如下：

```
GaussianNB (with SVD): Accuracy=0.6784, Precision=0.8961, Recall=0.4036, F1=0.5565
```

热力图：



为了验证高斯朴素贝叶斯在文本情感分析中的失效，我们在 TF-IDF 特征基础上进行了 Truncated SVD 降维，使得输入满足连续稠密形式以适配 GaussianNB。然而，降维后模型性能显著下降，F1 值由约 0.85 降至 0.69。这表明问题并非源于输入稀疏性，而**在于高斯假设本身与文本特征分布的本质冲突**。SVD 虽生成了连续特征，但破坏了特征独立性并引入符号对称性，进一步放大了高斯假设的不适用性。说明其高斯分布假设与文本特征的稀疏非负性质不匹配，验证了该方法在情感分析中的失效。

总结：在本次实验中，高斯朴素贝叶斯模型在 IMDB 电影评论情感分析任务中的表现显著低于其他模型，其失效的根本原因在于特征分布假设不符。该模型假设各特征在类别条件下服从高斯分布，而文本的 TF-IDF 特征具有非负、稀疏且高度偏态的分布特征，难以满足正态性要求。即使通过 SVD 降维将特征转化为稠密连续形式，特征之间的相关性和符号对称性仍破坏了独立同分布假设，使得模型无法构建有效的决策边界。因此，高斯朴素贝叶斯不适用于此类高维稀疏文本特征的情感分类任务。

四. 实验总结

本次实验以斯坦福电影评论数据集为基础，系统完成了文本情感分析的建模与验证过程。通过 TF-IDF 特征提取，将文本数据成功转化为可用于机器学习的高维稀疏表示，并分别训练了 Logistic Regression、Linear SVM 和 Multinomial Naive Bayes 三种分类模型。实验结果表明，线性模型（尤其是 SVM）在情感分类任务中表现最佳，F1 值接近 0.90，而 MultinomialNB 在保持较高效率的同时也能取得较好效果。相比之下，高斯朴素贝叶斯的分类性能明显较低，其主要原因在于高斯分布假设与 TF-IDF 特征的稀疏非负特性不符。整体而言，本实验帮助我掌握了文本特征提取、模型训练、性能评估与误差分析的完整流程，并通过对比不同算法的结果，加深了对模型假设与数据特性匹配关系的理解。